ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

*Appl. Statist.* (2020)
**69**, *Part* 4, *pp.* 815–839

# Global household energy model: a multivariate hierarchical approach to estimating trends in the use of polluting and clean fuels for cooking

Oliver Stoner, Gavin Shaddick and Theo Economou

*University of Exeter, UK*

and Sophie Gumy, Jessica Lewis, Itzel Lucio, Giulia Ruggeri and Heather Adair-Rohani

*World Health Organization, Geneva, Switzerland*

[Received November 2019. Revised May 2020]

**Summary.** In 2017 an estimated 3 billion people used polluting fuels and technologies as their primary cooking solution, with 3.8 million deaths annually attributed to household exposure to the resulting fine particulate matter air pollution. Currently, health burdens are calculated by using aggregations of fuel types, e.g. solid fuels, as country level estimates of the use of specific fuel types, e.g. wood and charcoal, are unavailable. To expand the knowledge base about effects of household air pollution on health, we develop and implement a novel Bayesian hierarchical model, based on generalized Dirichlet–multinomial distributions, that jointly estimates non-linear trends in the use of eight key fuel types, overcoming several data-specific challenges including missing or combined fuel use values. We assess model fit by using within-sample predictive analysis and an out-of-sample prediction experiment to evaluate the model's forecasting performance.

*Keywords*: Air pollution; Bayesian hierarchical model; Forecasting; Generalized Dirichlet distribution; Households; Solid fuels

## 1. Introduction

In 2017, an estimated 3 billion people, or 39% of the global population, used a solid fuel (charcoal, coal, crop residues, dung or wood) or kerosene as their primary fuel for cooking. This results in the emission of dangerous levels of pollutants, including fine particulate matter, $PM_{2.5}$, and carbon monoxide (World Health Organization, 2014). The World Health Organization (WHO) has estimated that about 3.8 million deaths per year world wide can be attributed to pollution from household cooking (World Health Organization, 2018a). This harm is compounded by the burden on people—notably women and children, who must dedicate large amounts of time to fuel collection which might otherwise be spent on education or work—and the risk of burn injuries.

To address this leading cause of disease and premature death in low and middle income countries, the '2030 agenda for sustainable development', which has been adopted by all United Nations member states, set targets of universal access to clean fuels and technologies for cooking (sustainable development goal (SDG) 7.1.2) and to reduce substantially the number of deaths from the joint effects of ambient and household air pollution (SDG 3.9). Although there have

*Address for correspondence*: Oliver Stoner, Department of Mathematics, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK.
E-mail: O.R.Stoner@exeter.ac.uk

been improvements in the proportion with access to clean fuels and technologies in some regions, globally these have been largely outpaced by population growth. This means that the absolute number of people without access to clean fuels and technologies has stagnated, decreasing only by 3% between 2000 and 2017. As a result, the world is projected to achieve only 74% clean fuel use by 2030 under current policy scenarios (SDG 7 Custodial Agencies, 2019).

In 2016 the World Health Assembly adopted a roadmap consisting of four priority areas of action to tackle the health risks of air pollution, notably 'expanding the knowledge base about impacts of air pollution on health' (World Health Organization, 2016). Currently, the WHO publishes estimates of 'polluting fuel use' and 'clean fuel and technology use', representing the combined use of all polluting fuels and all clean fuels and technologies respectively, for SDG monitoring. Here 'use' is defined as the proportion of people primarily relying on a given fuel or technology for cooking. In addition, the WHO at present assumes that 'clean fuel use equals clean fuel and technology use', because of the limited availability of data on the types of stoves that are used for cooking and the current absence of any scalable biomass stoves which can be considered 'clean' for health. These estimates are available for most countries, separately for urban and rural areas where fuel use trends often differ systematically, and for each year between 1990 and 2017. Conventionally, these estimates then serve as a practical surrogate for estimating the global burden of disease that is associated with using polluting fuels for cooking (Bonjour *et al.*, 2013). However, basing estimates of health effects on the combined use of polluting fuels fails to take into account variation in the risks that are associated with different fuels and technologies. Recently, Shupler *et al.* (2018) introduced a method for estimating exposure for several specific types of fuel that takes into account variation in exposure between countries. Despite this, global burden-of-disease estimates based on the use of specific fuels remain unavailable, as this would also require global estimates of specific fuel use.

In this paper, by developing and implementing a novel model for the use of eight specific types of fuel, we make a substantial contribution to the expansion of the knowledge base on the effects of household air pollution. Our aims are

(a) to estimate trends in specific fuel usage, together with coherent estimates of uncertainty,
(b) to provide meaningful estimates of individual fuel usage for countries where data are limited and
(c) to predict present day fuel usage, addressing lags in data collection, and to project estimated trends into the future.

Trends in the use of specific types of fuel are modelled together with survey sampling variability, which may vary between urban and rural areas and by country. Where data for a given country are limited, the model structure can derive information from regional trends. The model allows for different fuel use trends in urban and rural areas and can produce predictions (with associated uncertainty) of future use of various types of fuel, providing policy makers with a baseline against which they can evaluate the effectiveness of future interventions.

The remainder of the paper is organized as follows: Section 2 provides details of the available data and the proposed modelling approach, including the implementation of a model using Markov chain Monte Carlo (MCMC) sampling; Section 3 presents posterior predictive model checking and a future forecasting experiment; finally, Section 4 provides an overall summary and a concluding discussion of the model's impact.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`.

## 2. Methodology

Information on the types of technologies and fuels that are used by households for cooking is regularly collected in nationally representative household surveys or censuses and compiled in the WHO household energy database (World Health Organization, 2018b). At mid-2019, the database contained over 1100 surveys, with over 150 countries having at least one survey over the period 1990–2017. For each survey, the database contains the proportion of surveyed households using as their primary cooking fuel each of 10 key types: biogas; charcoal; coal; crop residues; dung; electricity; kerosene; liquid petroleum gas (LPG); natural gas; wood.

Over the period 1990–2017, the average number of surveys per country per year was around 0.3. Even if survey coverage were far greater, survey sampling variability means that individual surveys would still not be a reliable indicator on which to base policy decisions. Statistical models can be used to separate trends from sampling variability, while also enabling uncertainty in the trends to be appropriately quantified. Information from other sources, such as economic or social indicators, can also be included to allow for more reliable inference in countries with few surveys. For example, Rehfuess *et al.* (2006) used regression methods to quantify the association between solid fuel usage and a number of socio-economic factors to predict usage in countries where no data were available. An alternative source of information which can be exploited by statistical models is that the proportion of people using each type of fuel as their primary cooking fuel tends to be more similar, on average, between countries in the same region, than between countries in different regions. Fig. 1 illustrates differences in wood use by WHO region, with smooth density estimates of the proportion of households using wood as their primary cooking fuel, from surveys in years from 1990 to 2010. For example, the density estimates suggest that the use of wood is more prevalent in African countries than in European countries over this period.

The main modelling approach behind WHO monitoring of worldwide clean cooking fuel use was that of Bonjour *et al.* (2013) in the years leading up to 2018. Trends in overall solid fuel use (and more recently polluting fuel use) were estimated by using a multilevel (mixed effects) modelling approach. Unlike earlier regression-based approaches (Mehta *et al.*, 2006), Bonjour *et al.* (2013) did not include any covariates (e.g. national income). Instead, they relied exclusively on regional structures and smooth functions of time to estimate fuel use. To the best of our knowledge, the work that we present here constitutes the first major effort to estimate trends in the use of specific fuels for cooking. With that in mind, using data from the WHO

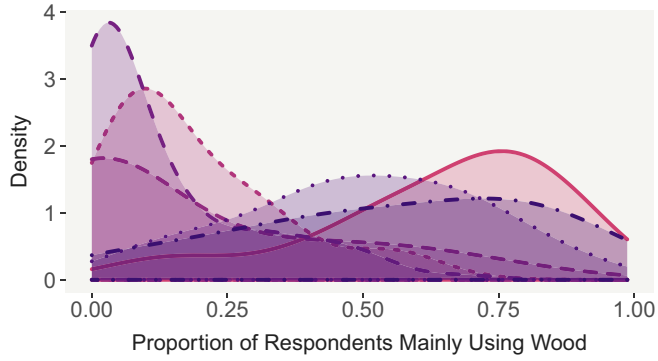


**Fig. 1.** Smooth density estimates of the proportion of survey respondents relying on wood as their primary cooking fuel by WHO region, from all surveys contained in the WHO household energy database (1990–2017) (source: WHO household energy database 2019): ———, Africa; - - - - -, Americas; — —, eastern Mediterranean; — —, Europe; · · · · , south-east Asia; · — · —, western Pacific

household energy database to estimate trends in specific fuel use presents some challenges that are related to inconsistencies in both the quality and the quantity of information that is available from the surveys. We specifically address four of these issues in our modelling approach.

(a) Many surveys report fuel values which are in some sense incomplete. This often includes combining more than one specific type of fuel (e.g. LPG and natural gas) into a single option in the survey (e.g. gas). In some cases this can arise because cultures and/or languages have a single term which includes several distinct types of fuel (e.g. the French language term 'charbon' which can include both coal and charcoal). Another common problem is inconsistency in how subfuels are categorized: for example, grass may be included in the crop residues category in one survey and in the dung category in another. Other, less common, issues include non-exhaustive lists of individual fuel options, with key fuels included in an 'other' category, resulting in missing values for those fuels. These issues mean that the time series of survey values for some fuels in some countries can be highly unstable.
(b) The total number of respondents is available for only approximately 50% of surveys in the database. For surveys where this information is not available, only the proportions using each fuel are given and the original counts (the number of respondents using each fuel) are non-recoverable.
(c) Information on trends in the use of specific fuels is required for both urban and rural areas but, in many cases, surveys provide data for only the overall population.

### 2.1. Generalized Dirichlet–multinomial model

For clarity of exposition, the following explanation relates to $y_i$, the number of respondents in a survey using fuel type $i$ as their primary fuel for cooking, ignoring for now any indices that are related to the country and the year. If we knew the total number of survey respondents $n$ for all data, a first approach to modelling could be to assume that data on $\mathbf{y} = \{y_i\}$ arise from a multinomial$(\mathbf{p}, n)$ distribution. Then $p_i$ would represent the proportion of people in the population using fuel $i$. This assumes that the survey sample is representative of the overall population. In reality, survey samples are imperfect and the multinomial model may not be sufficiently flexible to capture the extra variability that is caused by flaws in the survey design. For instance, the survey may not cover the whole geographical area of interest.

A flexible extension of this approach is to model $\mathbf{y}$ by using a generalized Dirichlet multinomial$(\boldsymbol{\alpha}, \boldsymbol{\beta}, n)$ (GDM) distribution (Zhang *et al.*, 2017): a mixture of the generalized Dirichlet (GD) model with probability density function

$$p(p_1, p_2, \ldots, p_k|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p_k^{\beta_{k-1}-1} \prod_{i=1}^{k-1} \left\{ \frac{p_i^{\alpha_i - 1}}{B(\alpha_i, \beta_i)} \left( \sum_{j=i}^{k} p_j \right)^{\beta_{i-1}-(\alpha_i+\beta_i)} \right\} \tag{1}$$

and the multinomial distribution, so that

$$\mathbf{p} \sim \text{generalized-Dirichlet}(\boldsymbol{\alpha}, \boldsymbol{\beta}); \qquad \mathbf{y}|\mathbf{p} \sim \text{multinomial}(\mathbf{p}, n). \tag{2}$$

The marginal probability mass function of the GDM is then

$$p(y_1, y_2, \ldots, y_k|\boldsymbol{\alpha}, \boldsymbol{\beta}, n) = \frac{\Gamma(n+1)}{\Gamma(y_k+1)} \prod_{i=1}^{k-1} \frac{\Gamma(y_i+\alpha_i)\Gamma\left( \sum_{j=i+1}^{k} y_j + \beta_i \right)}{B(\alpha_i, \beta_i)\Gamma(y_i+1)\Gamma\left( \alpha_i + \beta_i + \sum_{j=i}^{k} y_j \right)}. \tag{3}$$

Any additional variability that is caused by non-representative sampling can be potentially captured by the GD component. The GD model also has a very flexible covariance structure compared with the Dirichlet model (for example it can allow for positive covariance between elements of **p** (Wong, 1998)), which it reduces to in the special case that $\beta_i = \alpha_{i+1} + \beta_{i+1}$ for $i \in 1, \ldots, k-2$ and $\beta_{k-1} = \alpha_k$.

The GDM model has strong potential as a flexible regression framework for multivariate count data (Zhang *et al.*, 2017) that sum to a total. For instance, Stoner and Economou (2019) used the GDM to model reporting delays in time series of infectious disease counts. Nevertheless, the GDM model has seen little use in the modelling literature. Its use here is novel as a model for estimating trends and regional hierarchies in compositional data and for quantifying survey variability.

Recall from challenge (b) at the start of Section 2 that, for around half of the available data, only the proportion $\mathbf{x} = \{y_i/n\}$ of respondents using each fuel is available, with the total number of respondents $n$ being unknown. This means that we cannot use the GDM to model directly the number of respondents primarily using each fuel, if we wish to use all of the available data. However, as the principal interest lies in estimating or predicting trends in the fuel usage proportions **x**, an alternative approach would be to model the proportions themselves, e.g. by using a GD distribution. In that case, though, the presence of many 0% and 100% fuel usage observations (which fall outside the range space of the GD model) make this impractical. Instead, we opt for an approximate procedure for modelling **x**—which may also be applicable to other compositional data with a 0–1 inflation problem—where we transform observations of $x_i$ into conceptual counts $v_i$, out of a chosen total $N$. To ensure that the sum of the transformed counts does not exceed $N$, one can compute $v_i = \lfloor N x_i \rfloor$ (using the floor function, as opposed to rounding). The counts **v** can then be modelled as $\mathrm{GDM}(\boldsymbol{\alpha}, \boldsymbol{\beta}, N)$, so that predictions are based on $v_i/N$. The idea behind this is that the flexibility of the GDM model means that we can still capture the distribution of **x** well: any variability that is lost or gained from the multinomial component, by respectively using a larger or a smaller $N$ compared with the original $n$, can be accounted for by appropriate adjustment in the parameters of the GD component. From a modelling perspective, this amounts to estimating GD parameter values that adequately capture the total survey variability arising from finite and non-representative sampling. To utilize the original $n$ appropriately where they are known, we would need to estimate two separate sets of GD parameters: one set for when $n$ is known, to capture survey sampling variability minus that arising from finite sampling, and a second set for when $n$ is unknown, to capture all survey sampling variability. Alternatively, although not the approach that maximizes the available information in the data, treating all $n$ as unknown affords a significant reduction in the number of parameters to estimate. We therefore choose to treat all $n$ as unknown on the basis of practicality, subject to thorough model checking in Section 3 and Appendix C.

In Appendix A, we present a simulation study using the observed sample sizes $n$ from the data. Indeed, we illustrate that this approximate method (where all $n$ are assumed unavailable for modelling) yields an inference for the populationwide fuel usage which converges (as $N$ increases) to the inference that is obtained by modelling **y** directly. Our simulation experiment suggested that values greater than $N = 10000$ are probably sufficiently large, so we conservatively opt for $N = 100000$. This results in a virtually zero contribution to the variability of $\mathbf{v}/N$ from the multinomial component, bearing in mind that the GD component can absorb any additional variation that is associated with smaller sample sizes.

### 2.2. Tiered approach

To motivate the way in which we shall employ the GDM for these data, it is instructive to consider

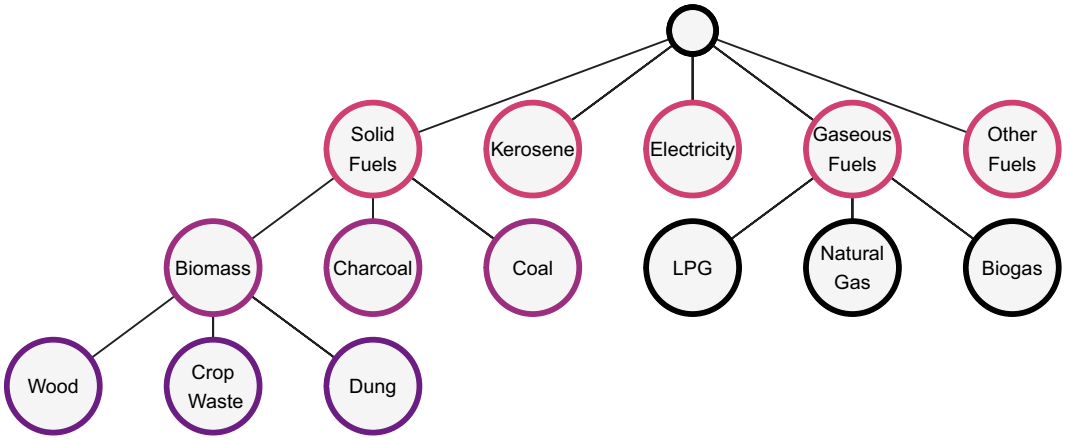


**Fig. 2.** Hierarchy of types of cooking fuel in the global household energy model

Fig. 2, which illustrates key types of cooking fuel and how they are typically aggregated into more general classifications, e.g. solid fuels. In principle, it is possible to model the use of specific fuels directly by using the GDM distribution:

$$v_1, \ldots, v_{11} \sim \mathrm{GDM}(\alpha, \beta, N); \tag{4}$$

$$\{1, \ldots, 11\} \equiv \{\text{wood, crop waste, dung, charcoal, coal, kerosene, electricity, LPG, natural gas, biogas, others}\}. \tag{5}$$

Predictions for aggregate groups, e.g. solid fuels, can then be achieved by aggregating predictions for the individual fuels. However, recall that one of the key challenges with modelling these data, challenge (a), is inconsistency in data collection. For example, some surveys combine more than one type of fuel (e.g. charcoal and coal) into a single category. Furthermore, there is sometimes inconsistency in the way that surveys categorize subfuels (e.g. grass). The result of this issue is that, for some countries, the time series of affected individual fuels are unstable. As such, modelling the use of all individual fuel types with one GDM distribution (as in expression (4)) will adversely impact estimates for the mean trends, sampling variability and any associated uncertainty, not just for affected fuels but for the other fuels as well, owing to the multivariate nature of the model and the data.

Fortunately, as they are the result of 'confusion' among certain types of fuel, these issues can be resolved by aggregating individual fuels into more general types of fuel. For example, confusion between wood, crop waste and dung can be resolved by aggregating data for these fuels into the more general category 'biomass' (which in this paper includes raw or unprocessed biomass fuels but excludes charcoal), and any outstanding confusion between charcoal and coal or between charcoal and wood can be resolved by aggregating into 'solid fuels'. Similarly, LPG and natural gas are very commonly combined at the survey level, which can be recognized by the formation of a 'gas' aggregate category.

This motivates the adoption of a tiered approach, where the use of the most aggregated fuel categories (e.g. solid fuels and gaseous fuels) are modelled as a GDM distribution at the 'top' tier (note that the tier does not relate to the merits or abundance of each fuel, only how we organize the fuels for modelling purposes), alongside other fuels that are unlikely to be confused or combined (e.g. kerosene and electricity) and an aggregation of other minor fuels and technologies (e.g. alcohol and solar stoves):

$$\{v_{\text{solid}}, v_{\text{kerosene}}, v_{\text{gas}}, v_{\text{electricity}}, v_{\text{others}}\} \sim \text{GDM}(\boldsymbol{\alpha}, \boldsymbol{\beta}, N). \tag{6}$$

This ensures that any instabilities arising from erroneous convolution of individual fuel types, e.g. charcoal and coal, does not propagate into the other fuel categories in the top tier. These categories can then be progressively disaggregated through nested GDM models. As in some countries there is a convolution between biomass fuel types (e.g. wood and crop waste), fully disaggregating solid fuels means that, in these countries, predictions for charcoal and coal will still be needlessly impacted. To address this, a 'mid'-tier is introduced to aggregate the biomass fuel types and to model these alongside charcoal and coal:

$$\{v_{\text{biomass}}, v_{\text{charcoal}}, v_{\text{coal}}\} \sim \text{GDM}(\boldsymbol{\alpha}, \boldsymbol{\beta}, v_{\text{solid}}). \tag{7}$$

The biomass fuel types can then be disaggregated in the 'lower' tier with a third GDM model:

$$\{v_{\text{wood}}, v_{\text{crop waste}}, v_{\text{dung}}\} \sim \text{GDM}(\boldsymbol{\alpha}, \boldsymbol{\beta}, v_{\text{biomass}}). \tag{8}$$

We could then disaggregate 'gas' into the three individual gaseous fuels with a fourth GDM model (a parallel mid-tier). This is, however, not essential for our application (estimating population exposure to household air pollution) as the difference between the different gaseous fuels in terms of pollutant concentrations is minimal compared with the difference between the gaseous fuels and the polluting fuels (World Health Organization, 2014). The upper, mid- and lower tiers are implemented together in a single Bayesian hierarchical model, so that uncertainty and variability are propagated both within and between tiers. Following this approach, the result is that a joint predictive inference for eight individual fuel types is achieved, but in a way which prevents inconsistency in particular types of fuel from affecting the others.

### *2.3. Conditional models*

Recall that an additional challenge, challenge (a) at the start of Section 2, is that occasionally a value $x_i$ (and thus $v_i$) is missing for at least one individual fuel (for a given country–year combination). To model these data in a way that easily allows prediction of the missing fuel values, we implement each GDM distribution (from the three tiers) by using the implicit conditional mass functions rather than the joint mass function. Specifically, for counts $\mathbf{v}$ and total $N$, the conditional distribution of (fuel) $v_i$ given the others is

$$v_i | v_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta} \sim \text{beta-binomial}\left(\alpha_i, \beta_i, n_i = N - \sum_{j<i} v_j\right), \tag{9}$$

$$p(v_i | v_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \binom{n_i}{v_i} \frac{B(v_i + \alpha_i, n_i - v_i + \beta_i)}{B(\alpha_i, \beta_i)} \tag{10}$$

Fitting this model in a Bayesian setting implies that any missing values $v_i$ can be sampled by using MCMC sampling. Furthermore, for ease of interpretation we reparameterize the conditional distributions in terms of their expectations $\nu_i$ and dispersion parameters $\phi_i$:

$$\begin{aligned} \alpha_i &= \nu_i \phi_i; \\ \beta_i &= (1 - \nu_i)\phi_i. \end{aligned} \tag{11}$$

The relative mean $\nu_i$ is interpreted as the expected proportion of households using fuel $i$ out of those not using any of the fuels higher up the hierarchy $(1, \ldots, i-1)$. For example, in the top tier GDM model $\nu_1$ is the expected proportion who use solid fuels from the whole population, $\nu_2$ is the proportion who use kerosene from the population who do not use solid fuels and $\nu_3$ is the proportion who use gas from the population who use neither solid fuels nor kerosene. Through

parameter $\phi_i$, the model can compensate for any gain or loss of variance in the conditional multinomial model for $v_i$ that is caused by the introduction of the 'artificial' total $N$. For more interpretable inference, the marginal mean vector of proportions $\boldsymbol{\mu} = \{\mu_i\}$ of households relying on each fuel $i$ can be recovered from the relative means $\nu_i$:

$$\begin{aligned} \mu_1 &= \nu_1; \\ \mu_k &= \nu_k \prod_{i=1}^{k-1} (1 - \nu_i) \qquad k \geqslant 2. \end{aligned} \tag{12}$$

### 2.4. Country and regional models

Introducing indices for a survey that is conducted in area $j$ (1, urban; 2, rural) of country $c$ and in year $t$, the characterization of the relative mean $\nu_{i,j,c,t}$ is defined by

$$\log\left(\frac{\nu_{i,j,c,t}}{1 - \nu_{i,j,c,t}}\right) = f_{i,j,c}(t), \tag{13}$$

where the logistic transformation ensures that $\nu_{i,j,c,t} \in (0, 1)$. Here we characterize functions $f$ as linear combinations of an intercept term, a linear term and non-linear thin plate spline terms:

$$f_{i,j,c}(t) = \xi_{0,i,j,c} + \xi_{1,i,j,c} X_{t,1} + \sum_{k=2}^{K} \xi_{k,i,j,c} X_{t,k}. \tag{14}$$

Here $\mathbf{X}$ is a model matrix of spline terms, where $X_{t,1}$ is linear in time and $X_{t,2}, \ldots, X_{t,K}$ are non-linear thin plate terms, and $\xi_{0,i,j,c}, \ldots, \xi_{K,i,j,c}$ are the corresponding coefficients. The choice of the number of basis terms $K$ must be made *a priori*, which corresponds to an upper bound on flexibility (similarly to choosing a number of polynomial terms). For larger $K$ the functions are penalized for smoothness parametrically. Here we choose $K = 10$: approximately one basis term for every 3 years, which we found was sufficiently high not to limit substantially the flexibility of the trends *a priori*. All coefficients are modelled as random effects, whose prior distributions have expectations (characterized as fixed effects) that vary with the region that the country is in (denoted by region index $r(c)$). Several choices are available for regional classifications, such as the six WHO regions, the 21 global burden of disease (GBD) regions and the seven GBD superregions (Shaddick *et al.*, 2018), or the eight SDG regions. For a chosen regional classification, the country random effects are modelled as

$$\xi_{0,i,j,c} \sim N(\gamma_{0,i,j,r(c)}, \sigma^{2(\xi)}_{0,i,j}), \tag{15}$$

$$\xi_{1,i,j,c} \sim N(\gamma_{1,i,j,r(c)}, \sigma^{2(\xi)}_{1,i,j}), \tag{16}$$

$$\{\xi_2, \ldots, \xi_K\} \sim \text{multivariate-normal}(\{\gamma_2, \ldots, \gamma_K\}, \boldsymbol{\Omega}^{-1(\xi)}_{i,j,c}). \tag{17}$$

The regional parameters (e.g. $\gamma_{0,i,j,r(c)}$) are then modelled as thin plate spline (fixed effect) coefficients. It can be shown that this model is equivalent to the additive combination of a regional level spline and a country level spline, where the former is a mean trend whereas the latter captures country level deviation from this mean. The advantage of this characterization over explicitly separating the country and regional effects into two different splines, however, is improved MCMC efficiency. Each precision matrix $\boldsymbol{\Omega}^{-1(\xi)}_{i,j,c}$ is known (Wood, 2016), and scaled by parameter $\lambda^{(\xi)}_{i,j,c}$ which penalizes the thin plate spline (specifically the deviation of the country spline from the regional spline) for smoothness, to avoid overfitting. Each $\log(\lambda^{(\xi)}_{i,j,c})$ is modelled as a random effect arising from an $N(\eta^{(\xi)}_{i,j}, \sigma^{2(\xi)}_{2,i,j})$ prior distribution.

The purpose of treating the coefficients $\boldsymbol{\xi}_{i,j,c}$ and smoothing parameters $\lambda^{(\xi)}_{i,j,c}$ as random effects is to improve prediction in countries with sparse data, where regional trends and global hyperparameters can constrain the overall country effect to be not too extreme with respect to other countries in the same region. We trialled this approach using the six WHO regions and alternatively the 21 GBD regions. Using the six WHO regions, we found that they encompassed too broad a range of fuel use patterns to be particularly useful for improving prediction in countries with little data. Conversely, when using the 21 GBD regions we found that they often contained too few countries, or had too many countries with little data, to allow the precise estimation of regional trends.

To address the issues that were posed by this choice, we opted for a nested model which utilizes both GBD regional structures and GBD superregional structures (denoted by index $s(r)$):

$$\gamma_{0,i,j,r} \sim N(\theta_{0,i,j,s(r)}, \sigma^{2(\gamma)}_{0,i,j}), \tag{18}$$

$$\gamma_{1,i,j,r} \sim N(\theta_{1,i,j,s(r)}, \sigma^{2(\gamma)}_{1,i,j}), \tag{19}$$

$$\{\gamma_2, \ldots, \gamma_K\} \sim \text{multivariate-normal}(\{\theta_2, \ldots, \theta_K\}, \boldsymbol{\Omega}^{-1(\gamma)}_{i,j,r}). \tag{20}$$

The regional thin plate spline coefficients (e.g. $\gamma_{0,i,j,r(c)}$) are now also modelled as random effects, with superregional expectations. Each precision matrix $\boldsymbol{\Omega}^{-1(\gamma)}_{i,j,r}$ is, as before, a known matrix scaled by parameter $\lambda^{(\gamma)}_{i,j,r}$, to penalize for smoothness (of the deviation of the regional trend from the superregional trend). The penalty parameters $\lambda^{(\gamma)}_{i,j,r}$ are once more modelled at the log-scale as random effects, arising from an $N(\eta^{(\gamma)}_{i,j}, \sigma^{2(\gamma)}_{2,i,j})$ prior distribution. The superregional intercept and linear terms are then modelled as fixed effects:

$$\theta_{0,i,j,s} \sim N(0, 10^2); \tag{21}$$

$$\theta_{1,i,j,s} \sim N(0, 10^2); \tag{22}$$

$$\{\theta_2, \ldots, \theta_K\} \sim \text{multivariate-normal}(\mathbf{0}, \boldsymbol{\Omega}^{-1(\theta)}_{i,j,s}). \tag{23}$$

Each $\boldsymbol{\Omega}^{-1(\theta)}_{i,j,s}$ is scaled by the fixed effect $\lambda^{(\theta)}_{i,j,s}$, whose prior distribution is $N(0, 10^2)$ at the log-scale. Now each $f_{i,j,c}(t)$ is equivalent to the additive combination of a superregional spline, a regional deviation spline and a country deviation spline. By adopting a nested regional structure, countries with little data benefit from more precise regional trend estimation, borrowing information from other countries in the same GBD region that have sufficient data. Failing that, further borrowing is achieved from the superregional trend.

Having specified models for the relative mean proportions $\nu_{i,j,c,t}$, it remains to define a model for the dispersion parameters $\phi_{i,c}$. Recall that these parameters are intended to capture additional survey variability (compared with the multinomial model), which is affected by the introduction of an 'artificial' sample size $N$. In countries with little data, we would like to constrain survey variability to reasonable levels, so we choose $\phi_{i,c}$ as random effects which vary by country:

$$\log(\phi_{i,c}) \sim N(\eta^{(\phi)}_{i,j}, \sigma^{2(\phi)}_{i,j}). \tag{24}$$

In the absence of any belief that survey variability should be regionally structured, the random effects are constrained by global hyperparameters $\eta^{(\phi)}_{i,j}$ and $\sigma^{2(\phi)}_{i,j}$.

### 2.5. Urban and rural variability

A further challenge with modelling these data, challenge (c) at the start of Section 2, is that, although most surveys in the data report both urban and rural values, some report only an overall value for the whole sample. So that these surveys can inform the estimation of urban and rural trends, we incorporate a layer in the model which relates the marginal mean proportions of urban, rural and overall values as follows:

$$\boldsymbol{\mu}_{c,t}^{\text{overall}} = \pi_{c,t}\boldsymbol{\mu}_{c,t}^{\text{urban}} + (1-\pi_{c,t})\boldsymbol{\mu}_{c,t}^{\text{rural}}; \tag{25}$$

$$\log\left(\frac{\pi_{c,t}}{1-\pi_{c,t}}\right) = \log\left(\frac{P_{c,t}}{1-P_{c,t}}\right) + g_c(t). \tag{26}$$

The overall mean fuel usage proportions $\boldsymbol{\mu}_{c,t}^{\text{overall}}$ are then defined as a weighted sum in equation (25), of the mean rural and urban proportions. The weights $\pi_{c,t} \in (0,1)$ represent the mean proportion of survey respondents living in an urban area, in country $c$ and year $t$. To capture structured demographic variability between countries and over time, United Nations (UN) estimates (United Nations, 2018) of the proportion of people living in an urban area for each country and year, $P_{c,t}$, are included as offsets in the model for $\pi_{c,t}$. For each country, any remaining structured variability in the urban proportion is modelled by using a smooth function $g_c(t)$. These functions should ideally be sufficiently flexible to capture the mean urban proportions well. However, from a modelling perspective, they also introduce extra degrees of freedom to capture the overall survey observations well. Therefore, to avoid overfitting, we once again employ penalized thin plate splines (with coefficients $\kappa_{0,c}, \ldots, \kappa_{K,c}$) for $g_c(t)$:

$$g_c(t) = \kappa_{0,c} + \kappa_{1,c}X_{t,1} + \sum_{k=2}^{K} \kappa_{k,c}X_{t,k}; \tag{27}$$

$$\kappa_{0,c} \sim N(0, \sigma_0^{2(\kappa)}); \tag{28}$$

$$\kappa_{1,c} \sim N(0, \sigma_1^{2(\kappa)}); \tag{29}$$

$$\{\kappa_2, \ldots, \kappa_k\} \sim \text{multivariate-normal}(\mathbf{0}, \boldsymbol{\Omega}_c^{-1(\kappa)}). \tag{30}$$

Each precision matrix $\boldsymbol{\Omega}_c^{-1(\kappa)}$ is, for one final time, a known matrix, scaled by a penalty parameter $\lambda_c^{(\kappa)}$ for smoothness. Then, $\log(\lambda_c^{(\kappa)}) \sim N(\eta^{(\kappa)}, \sigma_2^{2(\kappa)})$. Unlike the splines for $\nu_{i,j,c,t}$, the prior expectations are zero, as opposed to regional or superregional. This is because we have no prior belief that residual deviation from UN estimates in the sampling of urban respondents should be regionally structured. Employing thin plate splines here enables $g_c(t)$ to capture non-linear deviations from $P_{c,t}$ over time, but only when there is ample evidence in the data for a given country.

### 2.6. Robustness to outliers

In addition to the main data-specific modelling challenges that were highlighted in Section 2.1, the database contains some recorded values which truly defy the observed trend in their country. These values often cannot be explained by normal survey variability alone and can have an undue influence on the estimated trend if treated like ordinary observations. Although the beta–binomial conditional models that we employ are already more robust to outliers than equivalent binomial models, severe outliers can still cause issues, including causing the estimated trend to deviate substantially from other surveys to be closer to the outlier, or the overestimation of survey variability.

To address this problem, we model each observation as arising from a mixture distribution, which combines the beta–binomial conditional model with a discrete uniform distribution. The extent to which the model is either beta–binomial or uniform is controlled by the mixing parameter $\rho$. As $\rho$ approaches 0, the mixture becomes beta–binomial and vice versa:

$$p(v_i|v_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \rho \binom{n_i}{v_i} \frac{B(v_i + \alpha_i, n_i - v_i + \beta_i)}{B(\alpha_i, \beta_i)} + (1 - \rho)\frac{1}{n_i}. \tag{31}$$

This approach effectively enables the model to decide, given sufficient evidence in the data, whether or not a survey observation could plausibly have arisen from the same model as other nearby (in time) surveys for that country and area. The degree of evidence that is required can be controlled through the prior distribution that is specified for each $\rho$. For example, a strong prior distribution with most of the probability mass close to 0 for each $\rho$ corresponds to a strong belief that each survey value is very unlikely to be an outlier.

For this application, we introduce one $\rho$ for each unique survey. This means that, if a survey has an urban, rural and an overall value, a single $\rho$ controls the extent of mixing for all three. The reason for this is that if, for example, the model indicates that an urban value is a very severe outlier, we would prefer also to reduce the effect of the corresponding rural value on estimated trends and uncertainty. Including this layer in the model means that estimated trends are considerably more robust to outliers, as we shall highlight in Section 3.1. Additionally, predictions for $\rho$ are useful as an indicator to flag surveys efficiently that may warrant further investigation.

### 2.7. Prior distributions and implementation

For all hyperparameters $\eta$ which are the mean of a normal distribution (e.g. $\eta_{i,j}^{(\xi)}$), we specified non-informative $N(0, 10^2)$ prior distributions. For all hyperparameters $\sigma$ which are the standard deviation of a normal distribution (e.g. $\sigma_{0,i,j}^{(\xi)}$), we specified non-informative positive-truncated $N(0, 10^2)$ prior distributions.

All code was written and executed by using R (R Core Team, 2018) and the model was implemented by using NIMBLE (de Valpine *et al.*, 2017), which is a facility for highly flexible implementation of MCMC models. For this application, we needed to add the beta–binomial distribution to NIMBLE, which was straightforward with only a few lines of R code. Four MCMC chains were run for 80000 iterations from different randomly generated initial values and with different random-number generator seeds. The first 40000 samples were discarded as burn-in and, to limit system memory usage, the remaining samples were thinned by 10. Convergence of the MCMC chains is discussed in Appendix B. The model was applied to a subset of the data consisting of 1084 surveys and predictions were made for all countries with at least one survey (after selection). Survey selection criteria are discussed in Appendix D. Associated NIMBLE model code is available from `https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets` and data are available on request.

### 2.8. Prediction

The posterior predictive distribution for a new set of fuel use proportions in area $j$ (1, urban; 2, rural; 3, overall), $\mathbf{x}'_{j,c,t}$, given all of the observed fuel use proportions $\mathbf{x}$, is given by

$$p(\mathbf{x}'_{j,c,t}|\mathbf{x}) = \int\int p(\mathbf{x}'_{j,c,t}|\boldsymbol{\mu}_{j,c,t}, \phi_c)\, p(\boldsymbol{\mu}_{j,c,t}, \phi_c|\mathbf{x}) \mathrm{d}\boldsymbol{\mu}_{j,c,t} \mathrm{d}\phi_c. \tag{32}$$

Here $p(\mathbf{x}'_{j,c,t}|\boldsymbol{\mu}_{j,c,t}, \phi_c)$ represents the GDM likelihood for counts $\mathbf{v}'_{j,c,t}$, where $\mathbf{x}'_{j,c,t} = \mathbf{v}'_{j,c,t}/N$,

given mean fuel use $\boldsymbol{\mu}_{j,c,t}$ and variance parameters $\phi_c$. Also, $p(\boldsymbol{\mu}_{j,c,t}, \phi_c|\mathbf{x})$ represents the joint posterior distribution for $\boldsymbol{\mu}_{j,c,t}$ and $\phi_c$, given the data $\mathbf{x}$. Parameters $\boldsymbol{\mu}_{j,c,t}$ and $\phi_c$ are composed of country, regional and superregional effects, meaning fuel use predictions may be correlated between countries in the same region or superregion. The joint posterior $p(\boldsymbol{\mu}_{j,c,t}, \phi_c|\mathbf{x})$ is available to us in the form of MCMC samples, meaning that equation (32) can be evaluated by using Monte Carlo simulation. Specifically, for each MCMC sample of $\boldsymbol{\mu}_{j,c,t}$ and $\phi_c$ we can simulate a new vector of fuel use counts $\mathbf{v}'_{j,c,t}$ from the GDM with mean $\boldsymbol{\mu}_{j,c,t}$, variance parameters $\phi_c$ and total $N$, before calculating $\mathbf{x}'_{j,c,t} = \mathbf{v}'_{j,c,t}/N$. This results in a set of samples of $\mathbf{x}'_{j,c,t}|\mathbf{x}$, which is a Monte Carlo approximation of the posterior predictive distribution.

## 3. Model checking

The task of assessing the validity of the statistical model is divided into two parts: basic procedures to check that there are no systematic issues with reproducing the observed data and a forecasting experiment to evaluate the ability of the model to predict future fuel usage values.

### *3.1. Posterior predictive checking*

Given the Bayesian implementation of the model, assessing the fit to both in-sample and out-of-sample data is based on posterior predictive model checking (Gelman *et al.*, 2014). For in-sample data, this involves simulating replicates $\tilde{\mathbf{x}}|\mathbf{x}$ of the observed fuel proportions $\mathbf{x}$ from the posterior predictive distribution, using the approach that was described in Section 2.8. The statistical properties of these replicates can then be compared with properties of the corresponding observations. For brevity, we present predictive checking for solid fuel use in this subsection and for all the other fuel types in Appendix C.

In the first instance, scatter plots comparing the posterior means of the replicates with the observed values can give an indication of any systematic issues. These are shown for solid fuels in Fig. 3 and, for the most part, there are no obvious systematic problems. Also shown are coverage values: the proportion of observed solid fuel use values which lie within the 95% posterior predictive intervals, computed from the corresponding replicates. A coverage that is substantially lower than 95% would mean that a high proportion of observed values are extreme values with respect to the posterior predictive model, implying a poor fit. In this case, the coverage values for the 95% credible intervals were higher than 95% for all fuels and areas. Taken together, these two checks indicate that the model captures the observed data well.

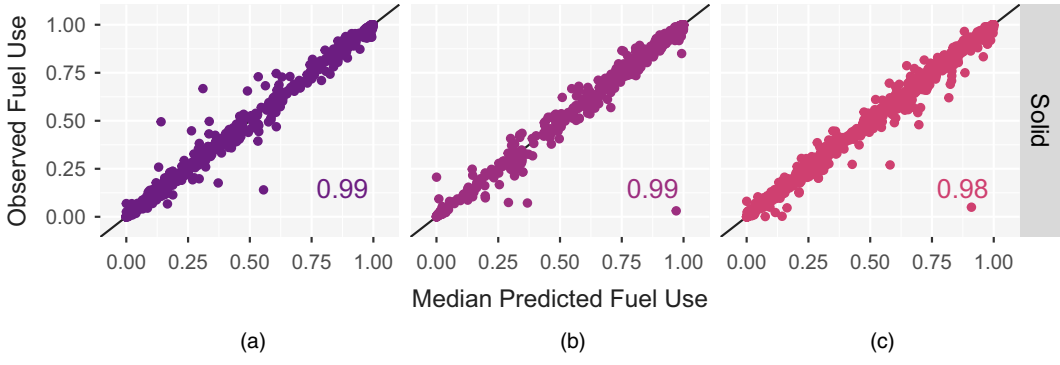


**Fig. 3.** Scatter plots comparing posterior means of solid fuel usage replicates $\tilde{x}_{1,j,c,t}$ with their corresponding observed values: (a) urban; (b) rural; (c) overall

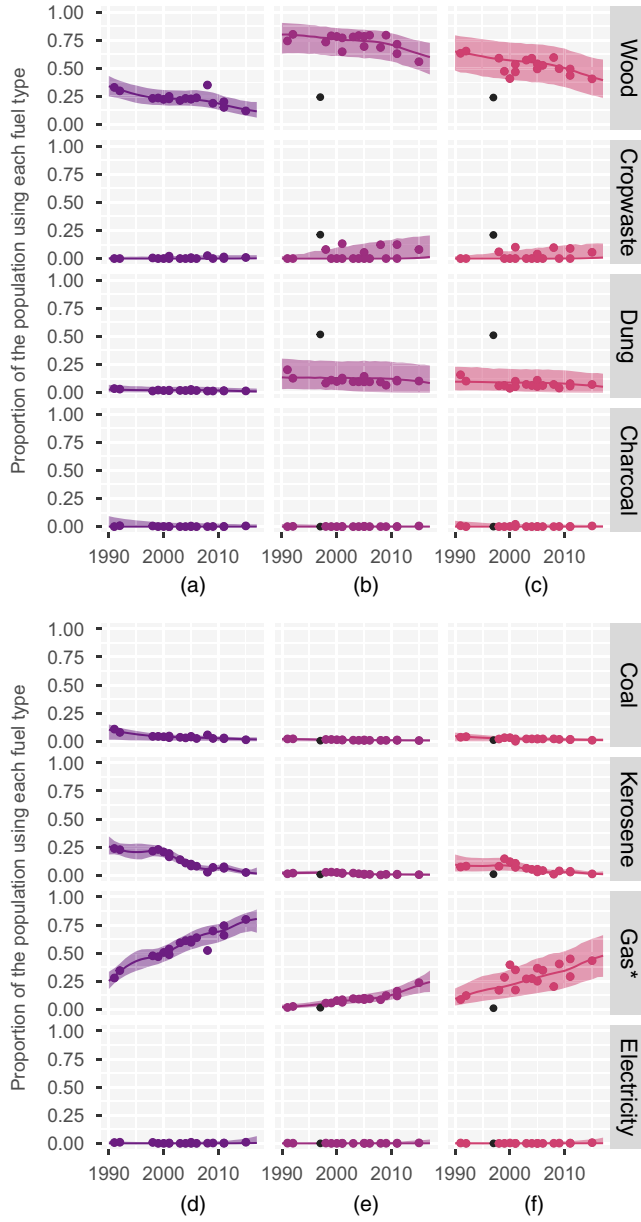


**Fig. 4.** Predicted fuel usage trends (median and 95% prediction intervals) for India (●, ●, ●, survey observations; ●, removed surveys; ⁎, LPG, natural gas and biogas): (a), (d) urban; (b), (e) rural; (c), (f) overall

Another way of checking the model is to compare predicted trends with survey observations on an individual country basis. Fig. 4 shows the median predicted proportion by using each fuel in each segment (urban, rural and overall) of India, with associated 95% posterior predictive intervals. Here it can be seen that the predicted trends follow the observed trends well, with prediction intervals that envelop a reasonable number of surveys. Moreover, by examining the tightness of the prediction intervals with respect to the variance of the observations, we can

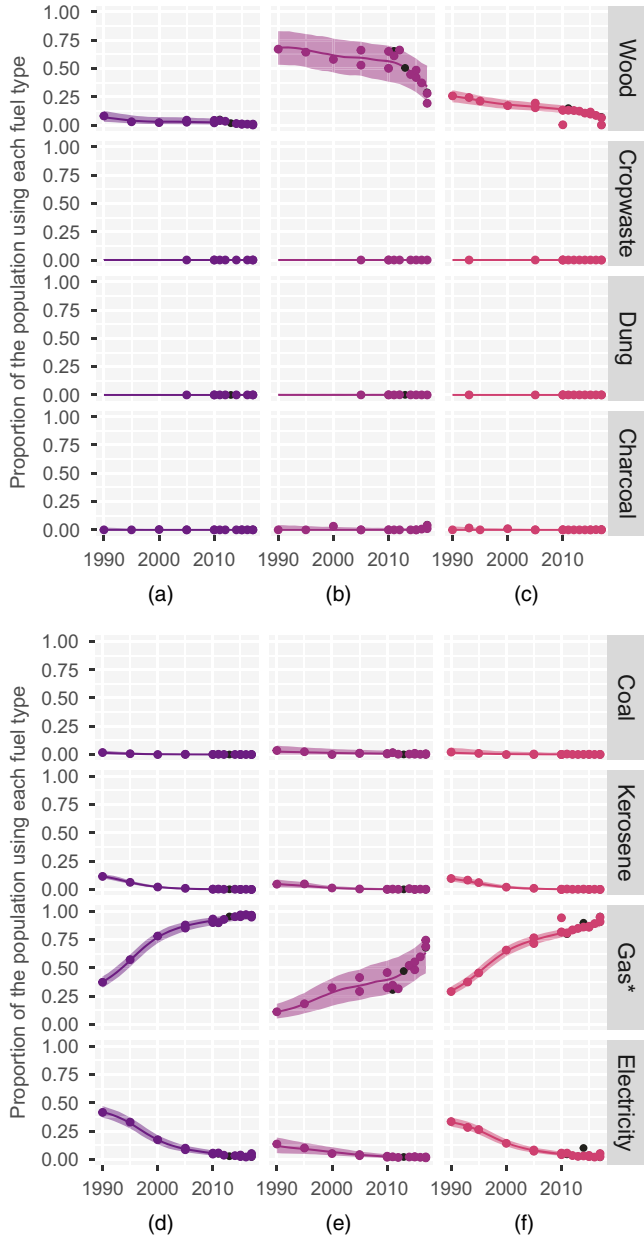


**Fig. 5.** Predicted fuel usage trends (median and 95% prediction intervals) for Colombia (●, ●, ●, survey observations; ●, removed surveys; ⁎, LPG, natural gas and biogas): (a), (d) urban; (b), (e) rural; (c), (f) overall

see that the high coverage values that are obtained for the replicate prediction intervals are not simply caused by excessively high model uncertainty.

A similar plot is shown for Colombia in Fig. 5. Looking at the use of gas in 2010, we can see that there is one survey with an unusually high overall value (specifically, this survey reports only an overall value). Through the $\rho$ corresponding to this survey, the model suggests that this

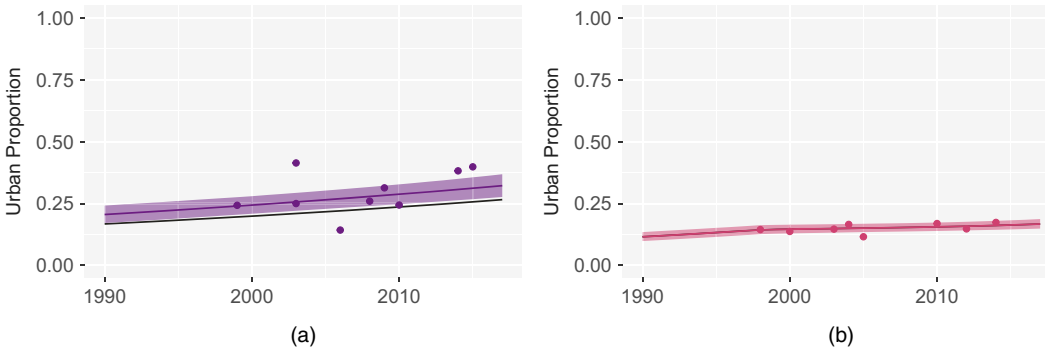


**Fig. 6.** Predicted mean survey urban proportions (———, ———) for (a) Kenya and (b) Malawi, compared with observed survey urban proportions (●, ●) and associated UN urban population estimates (———)

value is likely to be an outlier, such that the estimated trend and variability are not adversely affected. This illustrates the effectiveness of incorporating mixture distributions (as described in Section 2.6) in making the model more robust to outliers.

To check whether the model reproduces the observed data well, the overall predictions in Figs 4 and 5 incorporate the model's prediction of any systematic deviation $g_c(t)$ from the UN estimates of urban and rural proportions, in the sampling of urban and rural respondents. If desired, predictions of overall fuel usage can instead be based solely on the UN estimates of urban and rural proportions (rather than based on the proportions in the surveys). This is achieved by removing $g_c(t)$ from equation (26) during simulation.

Predicted fuel usage plots which include survey sampling variability (as in Figs 4 and 5) are included as on-line supplementary material for the eight most populous countries (at late 2019, excluding the USA and Russia).

We can also inspect the model's ability to capture structured between-country and temporal variability in the proportions of urban and rural respondents in the survey samples: Fig. 6 shows the proportion of (unweighted) respondents recorded as urban in the fuel surveys for Kenya (Fig. 6(a)) and Malawi (Fig. 6(b)) compared with UN estimates and predicted values from the model. The plot for Kenya shows evidence that the proportion of urban respondents in the surveys is, on average, higher than the UN estimates ($g_c(t) > 0$). The plot for Malawi, meanwhile, shows limited evidence of any systematic deviation ($g_c(t) \approx 0$). In both of these cases, the spline that is incorporated in equation (26) appears to capture any remaining structured variability (or the lack thereof) well, enabling reliable prediction of urban and rural trends where surveys provide values for only the overall population.

### *3.2. Forecasting experiment*

The model's ability to predict (forecast) fuel usage beyond the range of the data can be assessed by using out-of-sample predictive testing. This is important to validate the model's use for predicting present day fuel use, as there is a lag in data collection of 1–2 years, and for projecting estimated trends into the future, to provide a baseline against which the effects of interventions can be compared. To emulate a hypothetical forecasting scenario, the model was fitted only to surveys up to and including year 2012, therefore excluding 5 years (approximately 22% of the data). We then used the model to predict 5 years into the future and to produce predictive distributions for the out-of-sample surveys. As it is not our primary interest to forecast how any

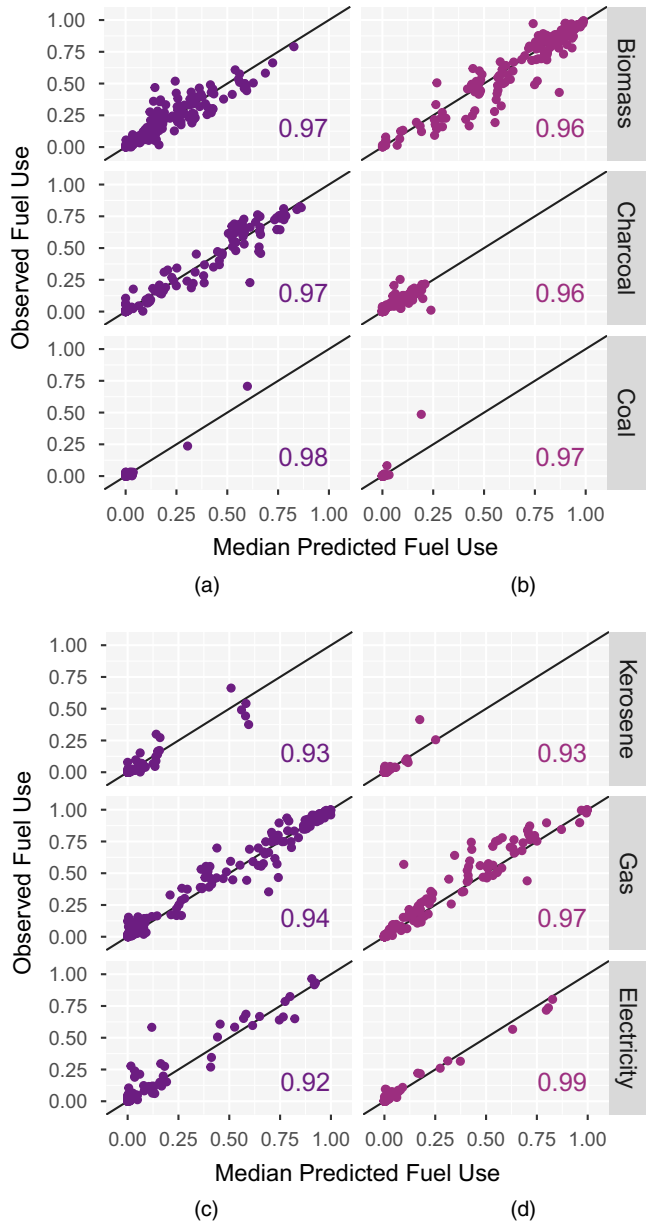


**Fig. 7.** Scatter plots of mean predicted fuel usage values from 2013 onwards, *versus* their observed values, from the model which was supplied data only from 2012 or earlier: (a), (c) urban; (b), (d) rural

systematic trends in the sampling of urban and rural respondents will progress in the future, we focus on checking the out-of-sample prediction of urban and rural surveys.

Fig. 7 shows scatter plots comparing the out-of-sample survey values with the mean predicted values from the model. Although there are some values which are not captured well (some potentially because of errors in data entry), generally the model does not seem to overpredict or underpredict systematically. Notably, the coverage values tend to be quite high, indicating that the model produces reliable uncertainty estimates when predicting into the future.

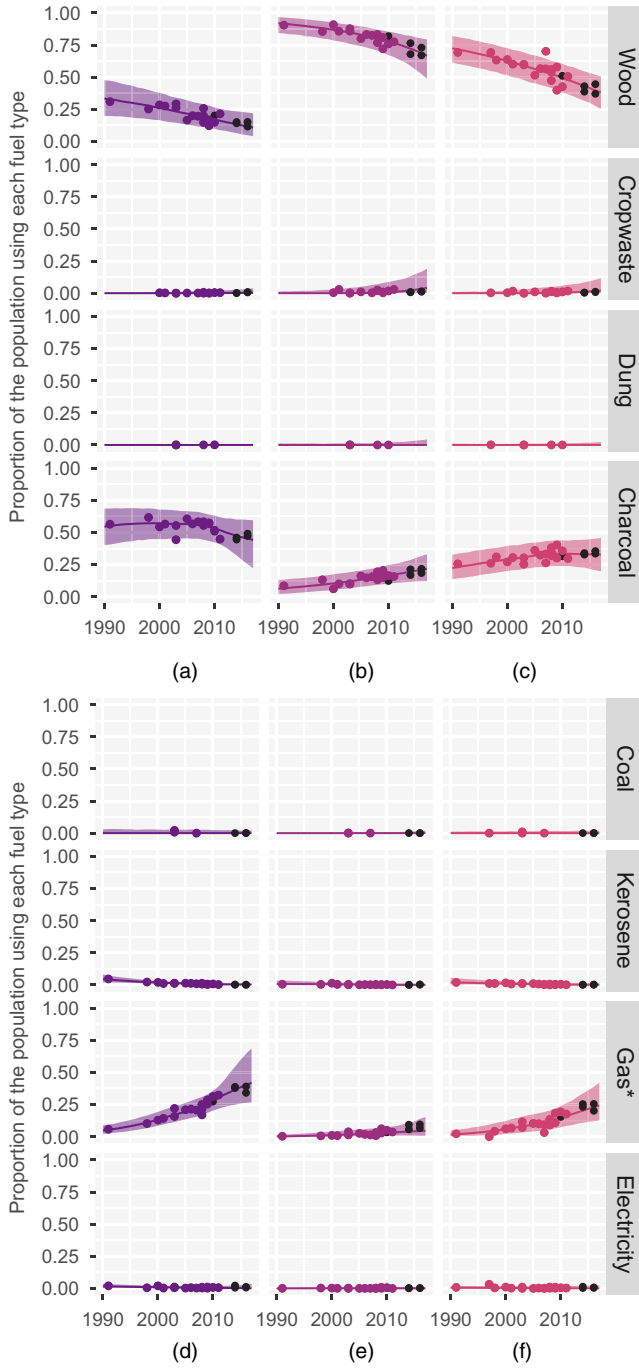


**Fig. 8.** Predicted fuel usage trends (median and 95% prediction intervals) for Ghana, from the model where surveys from 2013 onwards were excluded (the black points from 2013 onwards show excluded surveys) (⁎, LPG, natural gas and biogas): (a), (d) urban; (b), (e) rural; (c), (f) overall

To guard against high coverage values through unreasonably uncertain prediction intervals, we can assess the model's performance when forecasting by examining predictive plots for individual countries. Fig. 8 shows predictive fuel usage plots for Ghana, from the model where surveys from 2013 onwards are excluded. Here, the surveys removed are generally well within the 95% predictive intervals, which grow reasonably larger for predictions further into the future but are not so wide that they are impractical.

## 4. Discussion

Currently, the health burdens that are associated with exposure to air pollution from the use of polluting fuels for cooking are assessed on the basis of groupings of types of fuel (i.e. solid fuels or polluting fuels). However, this fails to take into account changes in the use of specific types of fuel that may affect the health impacts. For example, the results of the analyses that were performed here suggest that over the last few decades a substantial proportion of urban households in sub-Saharan Africa have switched from raw biomass fuels (i.e. wood, crop waste and dung) to charcoal, which has very different emissions characteristics. To expand the knowledge base about the effects of air pollution on health, burden-of-disease calculations should instead be based on the use of specific fuels, but until now country-specific estimates of specific fuel usage have been unavailable.

To address this, we have developed and implemented a novel multivariate hierarchical model for specific types of fuel which aims

(a) to estimate trends and associated measures of uncertainty, for specific fuels, for every country, and separately for urban and rural areas within a coherent modelling framework,
(b) to provide meaningful estimates in countries where there is limited data and
(c) to forecast fuel usage up to the present day and into the future.

Based on GDM distributions, the global household energy model automatically constrains the proportions of populations using each of eight key types of fuel ensuring that their sum does not exceed 1. Set within a Bayesian modelling framework, parametric and predictive uncertainty is quantified (e.g. by 95% prediction intervals) and verified by using within-sample posterior predictive checking (see Section 3.1). Where data availability is limited within a country, the model can 'borrow' information from neighbouring countries by using nested country, regional and superregional random effects, reducing predictive uncertainty. The model can forecast a number of years beyond the extent of the data, with assessment of forecasted values performed by using an out-of-sample predictive experiment (see Section 3). This allows present day fuel use to be evaluated, as data collection lags behind by 1–2 years. In addition, fuel use predictions for future years provide a baseline representation of what might be expected in the absence of intervention, with which future surveys that are conducted post interventions can be compared.

In achieving these aims, the model overcomes several challenges that are associated with using these survey data:

(a) inconsistency in survey design and collection, together with missing values, which can lead to highly unstable time series for some individual fuels in some countries;
(b) the total number of respondents is unavailable for around half of surveys;
(c) for many surveys, fuel use values are not available separately for urban and rural areas.

To address challenge (a), we adopted a tiered approach (Section 2.2) where we first modelled combined fuel use (e.g. solid fuels), which is progressively disaggregated into the component fuels. This ensures that excess variability and uncertainty among 'confused' fuels does not

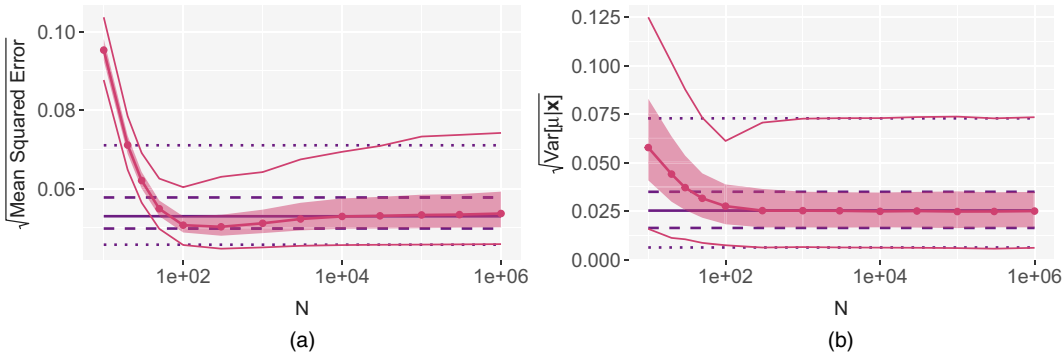


**Fig. 9.** (a) Median, interquartile range (■) and 95% interval (———) of the mean-squared differences between the posterior samples of the marginal mean proportions $\mu_{1,c},\ldots,\mu_{4,c}$ and their corresponding true values, from the approximate model with varying $N$ (●); (b) median, interquartile range (■) and 95% interval (———) of the posterior standard deviations of $\mu_{1,c},\ldots,\mu_{4,c}$): – – – –, results from the baseline model

propagate into those which are unaffected, and that predictions for the aggregate quantities are stable. To address the problem where the total number of respondents is unknown, challenge (b), we approximate a GDM model for the number of respondents, transforming the proportions by using each fuel into counts from an artificial sample size (Section 2.1). We illustrate that this results in approximately the same inference for populationwide fuel usage as modelling the (unavailable) number of respondents in their original count form, through a simulation experiment (which is presented in Appendix A). We addressed the unavailability of information on separate urban and rural use of fuel for all surveys, challenge (c), by including a layer in the model which links the urban, rural and overall fuel use values for each survey. Structured between-country and temporal variability in the proportion of urban respondents was then accounted for by combining UN estimates with smooth functions of time for each country. Finally, in addition to addressing these data-specific challenges, mixture distributions were employed to make the model more robust to potential outliers.

To date, the model has been adopted by the WHO to produce estimates of the proportion of people in each country who rely on polluting fuels as their primary fuel and technology for cooking and has played a central role in monitoring SDG 7.1.2 (SDG 7 Custodial Agencies, 2019). It has also played an important role in identifying data that appear to be out of line with general country level patterns for further investigation. Ultimately, the modelling approach proposed provides policy makers with decision quality information and enables a ground breaking reassessment of the health effects of cooking with polluting fuels and technologies.

## Acknowledgements

This work was supported by a Natural Environment Research Council GW4+ doctoral training partnership studentship (NE/L002434/1) and WHO contract APW 201790695.

## Appendix A: Simulation experiment

To illustrate the validity of our approximation for modelling the proportions using each type of fuel $\mathbf{x}=\mathbf{y}/n$, we present a simulation experiment using the 598 observed survey samples sizes $n$. The majority are in the range 1000–100000, with a mode of around 10000. At these large values, the contribution of the multinomial variance to the total variance of $\mathbf{x}$ would be small.

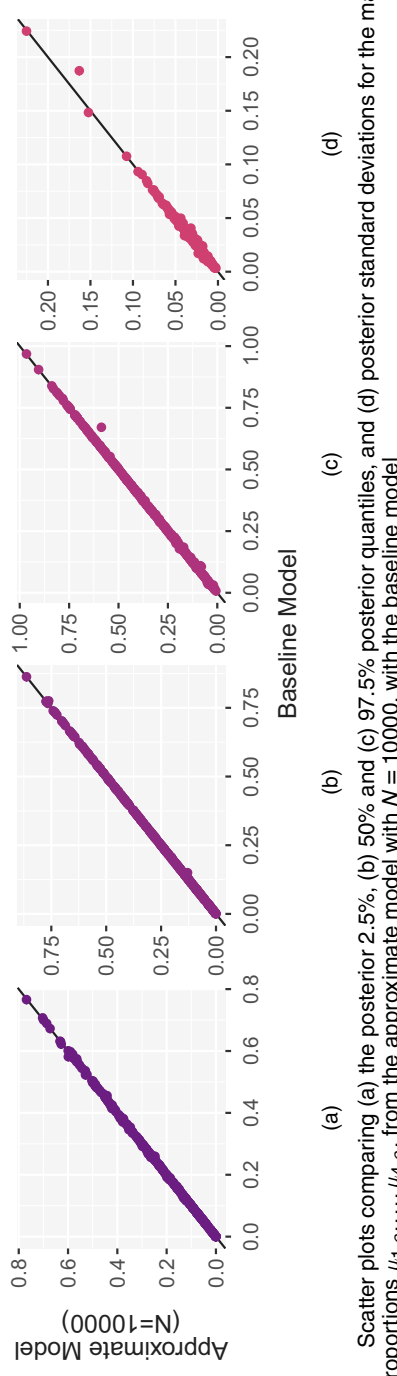


**Fig. 10.** Scatter plots comparing (a) the posterior 2.5%, (b) 50% and (c) 97.5% posterior quantiles, and (d) posterior standard deviations for the marginal mean proportions $\mu_{1,c}, \ldots, \mu_{4,c}$, from the approximate model with $N = 10000$, with the baseline model

For each available $n_i$ ($i = 1, \ldots, 598$), we simulate a vector of survey responses $\mathbf{y}_i = \{y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4}\}$ from a GDM model. Here, each country has a different (time constant) marginal mean vector $\boldsymbol{\mu}_c$ and variance parameters $\phi_c$ (preserving the original associations between the countries and observed $n_i$ in the data, and ignoring countries with no observed $n_i$). Some countries will have only one $\mathbf{y}_i$ and others will have several (each with its own unique $n_i$). We simulate all of the $\boldsymbol{\mu}_c$ from a Dirichlet($\mathbf{1}$) distribution, and all the $\phi_c$ independently from a gamma$(4, 0.1)$ distribution (inducing a moderately high degree of overdispersion, compared with the multinomial model):

$$\left.\begin{aligned} \boldsymbol{y}_i &\sim \text{GDM}(\boldsymbol{\mu}_c, \phi_c, n_i); \\ \boldsymbol{\mu}_c &\sim \text{Dirichlet}(\mathbf{1}); \\ \phi_c &\sim \text{gamma}(4, 0.1). \end{aligned}\right\} \tag{33}$$

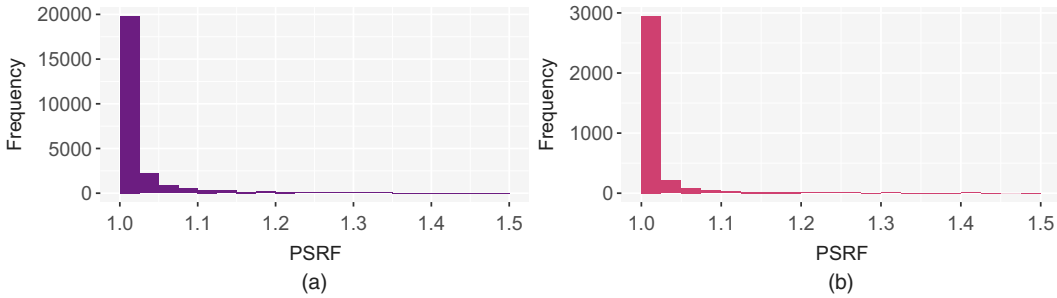


**Fig. 11.** Histograms of the PSRF for (a) the relative means $\nu_{i,j,c,t}$ and (b) variance parameters $\phi_{i,j,c}$

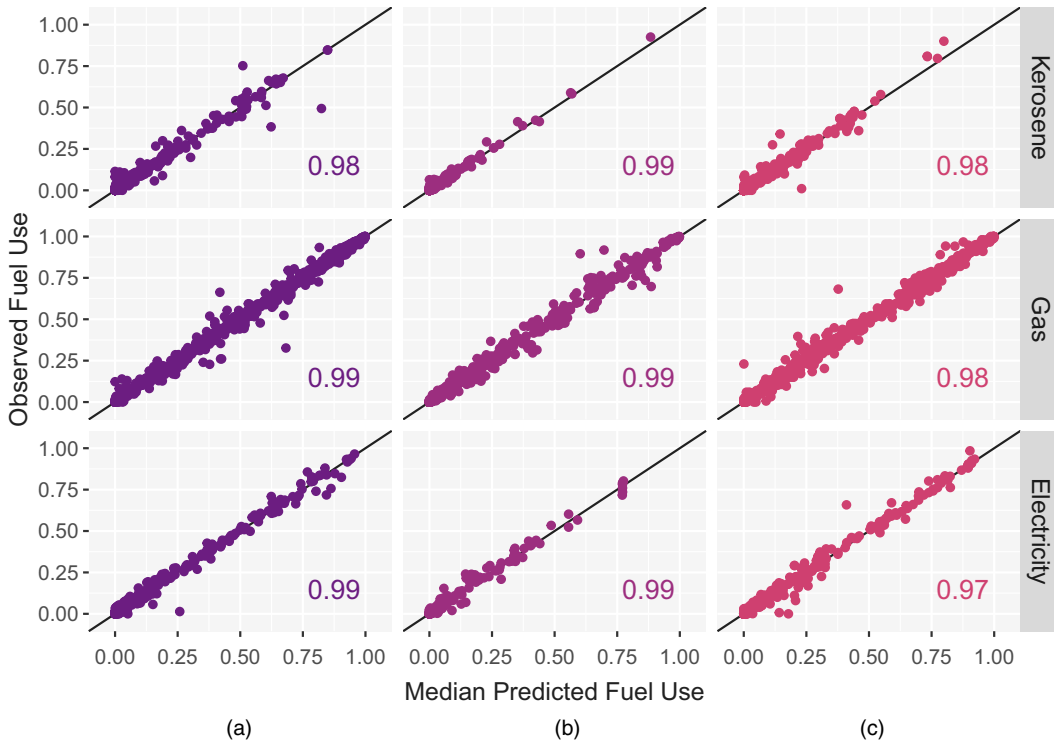


**Fig. 12.** Scatter plots comparing the posterior means of kerosene, gas and electricity use replicates with their corresponding observed values: (a) urban; (b) rural; (c) overall

In the baseline scenario, against which we shall compare our approximate method, we have observations for all the $n_i$ and all the $\mathbf{y}_i$. This enables us to implement the above model directly, which we do in a Bayesian setting using a Dirichlet($\mathbf{1}$) prior for each $\boldsymbol{\mu}_c$ and a non-informative exponential(0.001) prior for each $\phi_c$.

In the second scenario, we do not know any of the $n_i$ or the $\mathbf{y}_i$, but we do have observations for $\mathbf{x}_i = \mathbf{y}_i / n_i$. In this scenario, we can apply our approximate method (from Section 2.1), where we fit the GDM to constructed counts $\mathbf{v}_i = \lfloor N\mathbf{x}_i \rfloor$. We proceed to apply this method while varying $N$ over a range of values (10, 20, 30, 50, 100, 300, 1000, 3000, 10000, 30000, 100000, 300000, 1000000), so that we can investigate the effect of this choice on parameter inference.

Recall that in our application we are primarily interested in correct inference for the marginal mean proportions $\boldsymbol{\mu}_c$ (the populationwide fuel use in each country), and we claimed that a sufficiently large choice of $N$ yields a parameter inference that is approximately the same as if we had modelled the $\mathbf{y}_i$ directly, along with the sample sizes $n_i$. To assess this, we begin by examining the models' accuracy when predicting the true marginal mean proportions $\boldsymbol{\mu}_c$. For each posterior sample, we can compute the mean-squared error between the predicted values of $\boldsymbol{\mu}_c$ and the true values. Fig. 9(a) shows the median of this statistic, for varying $N$, as well as the interquartile range (dark), and 95% prediction interval (light). Compared with the same statistics for the baseline model, which are shown as horizontal lines, we can see that the distribution of mean-squared errors for the approximate method does indeed converge to the baseline model as $N$ increases, from about $N = 10000$ onwards.

We can also examine how the approximate method quantifies uncertainty in $\boldsymbol{\mu}_c$. For each individual $\mu_{1,c}, \ldots, \mu_{4,c}$, we compute the standard deviation of the posterior samples. The median of these posterior samples are then shown for each $N$ in Fig. 9(b), once again alongside the interquartile range and 95% interval. The distribution of posterior standard deviations for the approximate method also converges to the baseline model, but it does so for a much lower $N$ (between 100 and 1000) than does the mean-squared error.

Finally, if we choose a single value of $N$, we can compare more closely the approximate method with the baseline model when estimating $\boldsymbol{\mu}_c$. Fig. 10 compares the 2.5%, 50% and 97.5% posterior quantiles

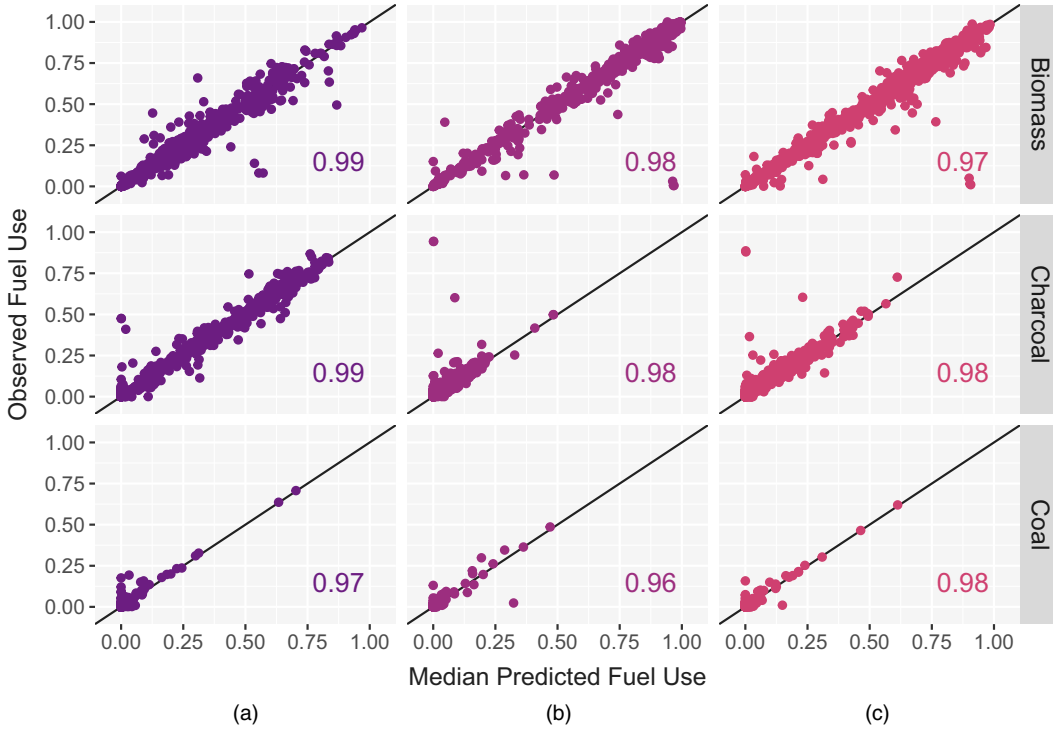


**Fig. 13.** Scatter plots comparing the posterior means of biomass, charcoal and coal use replicates with their corresponding observed values: (a) urban; (b) rural; (c) overall

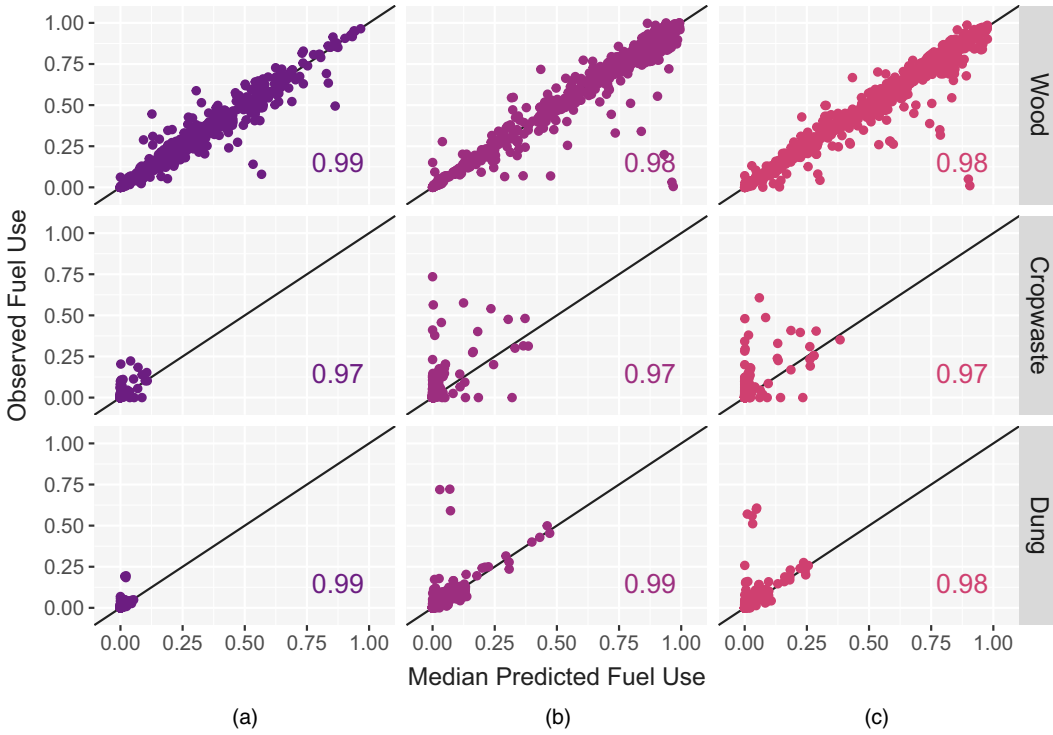


**Fig. 14.** Scatter plots comparing the posterior means of wood, crop waste and dung use replicates with their corresponding observed values: (a) urban; (b) rural; (c) overall

for the $\mu_{1,c}, \ldots, \mu_{4,c}$, from the approximate model with $N = 10000$, with the quantiles from the baseline model. The quantiles are virtually identical, suggesting that for these simulated data the same inference for $\boldsymbol{\mu}_c$ would be achieved either by modelling the true counts $\mathbf{y}_i$ directly or by modelling the constructed counts $\mathbf{v}_i = \lfloor 10000\mathbf{x}_i \rfloor$.

## Appendix B: Convergence of Markov chain Monte Carlo chains

One way to assess the convergence of MCMC chains is to compute the potential scale reduction factor (PSRF) for some key parameters. This compares the variance between the MCMC chains with the variance within the chains (Brooks and Gelman, 1998). A PSRF of 1 is obtained when the two variances are the same, so starting the chains from different initial values and obtaining a PSRF that is close to 1 (typically taken to be less than 1.05) gives a good indication that the chains have converged to the parameter's posterior distribution.

We computed the PSRF for the (26016) relative means $\nu_{i,j,c,t}$ corresponding to the survey observations and the (3576) variance parameters $\phi_{i,j,c}$. Figs 11(a) and 11(b) present these respectively in frequency histograms. For both sets of parameters, the overwhelming majority of the values lie in the closest bin to 1, suggesting that the model has converged.

## Appendix C: Further model checking

As discussed in Section 3.1, it is important to verify that the model can reproduce the observed data well. We do this by comparing replicates (predictions) of the observed data with the actual observations. In Section 3.1 we checked the replicates of solid fuel use and here we check the remaining fuels.

Fig. 12 shows scatter plots comparing the mean predicted replicates for the three other main top tier types of fuel, kerosene, gas and electricity, with their corresponding observed values. Similarly, Fig. 13

shows the same plots for the three mid-tier types of fuel, biomass, charcoal and coal, and Fig. 14 shows the three lower tier fuel types, wood, crop waste and dung. In general the points are scattered about the diagonal line fairly evenly, indicating a good model fit for the different fuels. Notably, however, the fit of the model is more precise for types of fuel in the upper tiers (e.g. electricity) than for those in the lower tier (e.g. dung). This makes sense, as these fuels are less likely to be affected by the issues that were described at the start of Section 2 and in Section 2.3, such as the combination of certain types of fuel, where some of the observed values are likely to be erroneous and difficult for the model to capture well. Regardless, the coverage of the 95% intervals is very high for all fuels.

## Appendix D: Survey selection

The model was applied to a selection of the WHO household energy database. Surveys were excluded from the analyses if they

(a) reported only the usage of 'solid fuels' as a group, rather than the usage of at least one individual fuel type,
(b) included an excessively high proportion (greater than 15%) of respondents who either reported that they cook with an unlisted fuel, that they do not cook at all or who failed to respond and
(c) were flagged in the database as unsuitable for modelling.

Surveys which were not included for modelling are shown as black points in the plots of predicted fuel use that are provided as on-line supplementary material.

Additionally, some surveys report household-weighted fuel use values (e.g. 50% of households use wood) whereas others report national population-weighted values (e.g. 50% of the population uses wood). Here we assume an equivalence between the two types of weighting, which is consistent with previous approaches to modelling fuel use (Bonjour *et al.*, 2013).

## References

Bonjour, S., Adair-Rohani, H., Wolf, J., Bruce, N. G., Mehta, S., Prüss-Ustün, A., Lahiff, M., Rehfuess, E. A., Mishra, V. and Smith, K. R. (2013) Solid fuel use for household cooking: country and regional estimates for 1980–2010. *Environ. Hlth Perspect.*, **121**, 784–790.

Brooks, S. P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *J. Computnl Graph. Statist.*, **7**, 434–455.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2014) *Bayesian Data Analysis*, 3rd edn. Boca Raton: Chapman and Hall–CRC.

Mehta, S., Gore, F., Prüss-Üstün, A., Rehfuess, E. and Smith, K. (2006) Modeling household solid fuel use towards reporting of the millennium development goal indicator. *En. Sustain. Devlpmnt*, **10**, 36–45.

R Core Team (2018) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rehfuess, E., Mehta, S. and Prüss-Ustün, A. (2006) Assessing household solid fuel use: multiple implications for the Millennium Development Goals. *Environ. Hlth Perspect.*, **3**, 373–378.

SDG 7 Custodial Agencies (2019) Tracking SDG 7: the energy progress report. *Report*. World Bank, Washington DC. (Available from `https://trackingsdg7.esmap.org/data/files/download-documents/2019-Tracking%20SDG7-Full%20Report.pdf`.)

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Van Dingenen, R., Dora, C., Gumy, S., Liu, Y., Martin, R., Waller, L. A., West, J., Zidek, J. V. and Prüss-Ustün, A. (2018) Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Appl. Statist.*, **67**, 231–253.

Shupler, M., Godwin, W., Frostad, J., Gustafson, P., Arku, R. E. and Brauer, M. (2018) Global estimation of exposure to fine particulate matter (PM 2.5) from household air pollution. *Environ. Int.*, **120**, 354–363.

Stoner, O. and Economou, T. (2019) Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, to be published.

United Nations (2018) World urbanization prospects: the 2018 revision. *Technical Report*. (Available from `https://population.un.org/wup/`.)

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T. and Bodik, R. (2017) Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J. Computnl Graph. Statist.*, **26**, 403–413.

Wong, T.-T. (1998) Generalized Dirichlet distribution in Bayesian analysis. *Appl. Math. Computn*, **97**, 165–181.

Wood, S. (2016) Just another Gibbs additive modeler: interfacing JAGS and mgcv. *J. Statist. Softwr.*, **75**, 1–15.

World Health Organization (2014) *WHO Guidelines for Indoor Air Quality: Household Fuel Combustion*. Geneva: World Health Organization.
World Health Organization (2016) Health and the environment: draft road map for an enhanced global response to the adverse health effects of air pollution: report by the secretariat. World Health Organization, Geneva. (Available from https://apps.who.int/iris/handle/10665/250653.)
World Heath Organization (2018a) WHO press release. World Health Organization, Geneva. (Available from https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action.)
World Health Organization (2018b) Household energy database. World Health Organization, Geneva. (Available from https://www.who.int/airpollution/data/household-energy-database/en/.)
Zhang, Y., Zhou, H., Zhou, J. and Sun, W. (2017) Regression models for multivariate count data. *J. Computnl Graph. Statist.*, **26**, 1–13.

*Supporting information*
Additional 'supporting information' may be found in the on-line version of this article.